ASSIGNMENT 2 HADOOP

ITRI 623

ENRICO
DREYER
31210783

Table of Contents

Introduction

For this assignment we were asked to make a video of no longer than 2 minutes on any two sections of the Hadoop Udemy Course. For the second Assignment I decided to go with using relational data stores with Hadoop.

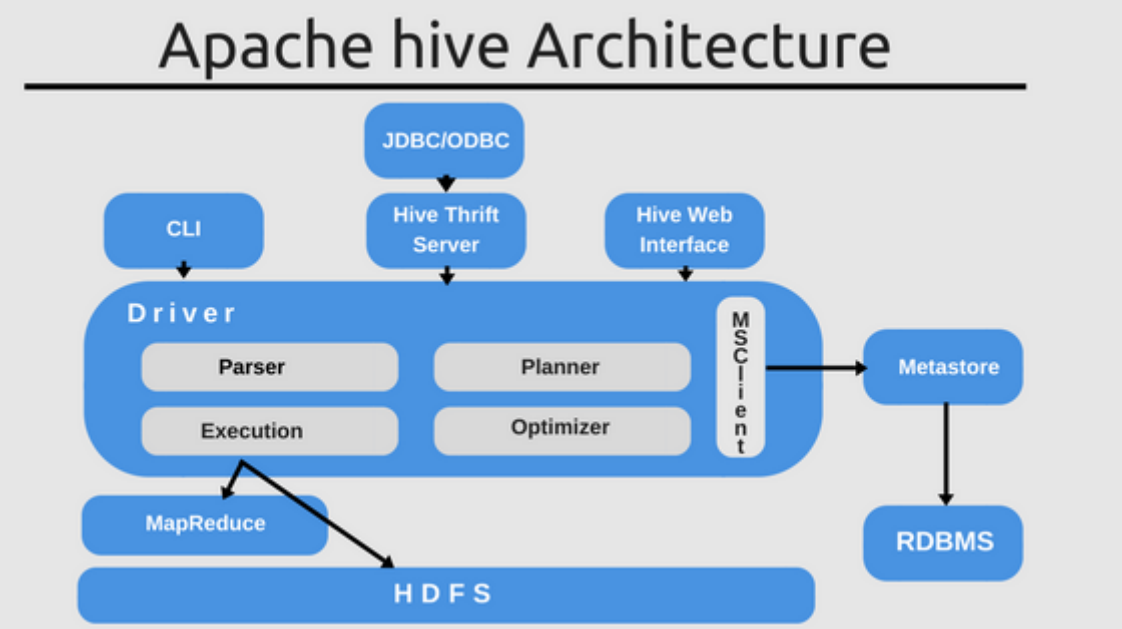
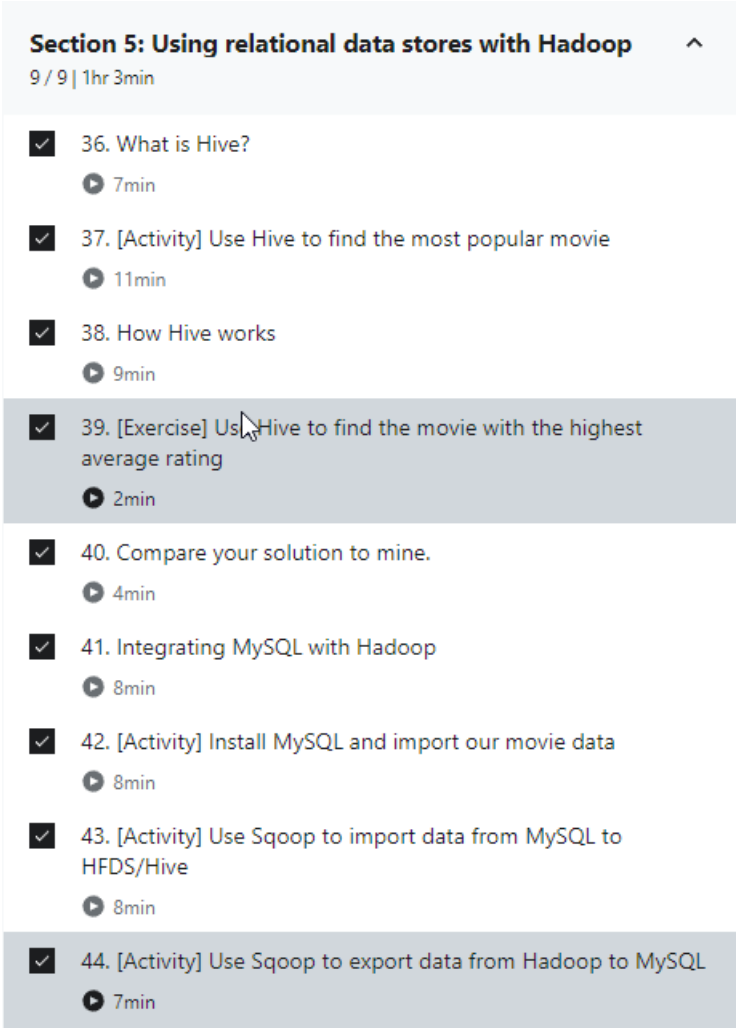


Figure 1: Hive architecture (Malhotra, 2018)

The following section of the document will discuss the notes of my video

Proof



My Random notes of the course:

What is Hive? - Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data and makes querying and analyzing easy (IBM, 2021).

OLAP Queries - Online analytical processing (OLAP) is a system for performing multi-dimensional analysis at high speeds on large volumes of data (Galaktikasoft, 2021).

Three kinds of OLAP types: ROLAP, MOLAP, HOLAP

You can upload your data via CSV!

Example query from course:

create view topmovieId's AS

select movieId, count(movieID) as ratingCount

From ratings

Group By movieID

Order By ratingCount Desc;

Google examples:

- **Rolling up** to country level:

```
SELECT COUNT(visits), SUM(sales)
GROUP BY country
```

Country	visits	sales
USA	4	$50
Canada	1	0

- **"Slice"** by browser

```
SELECT COUNT(visits), SUM(sales)
GROUP BY country
HAVING browser = "FF"
```

Country	visits	sales
USA	2	$10
Canada	0	0

- **Top** browsers by sales

```
SELECT SUM(sales), COUNT(visits)
GROUP BY browser
ORDER BY sales
```

Browser	sales	visits
Chrome	$25	2
Safari	$15	1
FF	$10	2

Figure 2: OLAP Examples (slideshare, 2013)

Apache Sqoop - Apache Sqoop is a command-line interface application used for transferring data between relational databases and Hadoop (IntelliPaat, 2021).

References

Galaktikasoft. (2021). OLAP and query language: How to write OLAP queries. https://galaktika-soft.com/blog/olap-essence-query-language.html

IBM. (2021). What is Apache Hive? https://www.ibm.com/analytics/hadoop/hive

IntelliPaat. (2021). What Is Apache Sqoop? https://intellipaat.com/blog/what-is-apache-sqoop/#:~:text=Sqoop%20lets%20you%20automate%20the%20process%2C%20and%20depending,and%20exporting%20data%2C%20providing%20a%20parallel%20fault-tolerant%20mechanism.

Malhotra, A. (2018). Apache Hive – A Faster and Better SQL on Hadoop. https://www.whizlabs.com/blog/apache-hive-faster-better-sql-on-hadoop/

slideshare. (2013). OLAP 101 – Queries example. https://www.slideshare.net/clehene/realtime-olap-for-big-data-use-cases-bigdataro-2013/6-OLAP_101_Queries_example_Rolling